Text to Matrix Generator Toolbox

A Brief Introduction

Eugenia Maria Kontopoulou, Dimitrios Zeimpekis
and Efstratios Gallopoulos

Department of
Computer Engineering and Informatics
University of Patras

Patras, 09/05/2014
1 Introduction to Text Data

2 The Text-to-Matrix Generator
## Documents

<table>
<thead>
<tr>
<th>Labels</th>
<th>Titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>Identifying users of social networks from their data footprint: An application of large-scale matrix factorizations</td>
</tr>
<tr>
<td>B2</td>
<td>Data fusion based on coupled matrix and tensor factorizations</td>
</tr>
<tr>
<td>B3</td>
<td>On incremental deterministic methods for dominant space estimation for large data sets</td>
</tr>
<tr>
<td>B4</td>
<td>Fast projection methods for robust separable nonnegative matrix factorization</td>
</tr>
<tr>
<td>B5</td>
<td>Experiments with randomized algorithms in the text to matrix generator toolbox</td>
</tr>
</tbody>
</table>
Text Data

... to Term-Document structures ...

Term-Document Matrix (TDM)

33 × 5

<table>
<thead>
<tr>
<th>terms</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>B5</th>
<th>terms</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>B5</th>
</tr>
</thead>
<tbody>
<tr>
<td>algorithm</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>matrix</td>
<td>0.3219</td>
<td>0.3219</td>
<td>0</td>
<td>0.3219</td>
<td>0.3219</td>
</tr>
<tr>
<td>applic</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>method</td>
<td>0</td>
<td>0</td>
<td>1.3219</td>
<td>1.3219</td>
<td>0</td>
</tr>
<tr>
<td>base</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>network</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>couple</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>nonneg</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
</tr>
<tr>
<td>data</td>
<td>0.7370</td>
<td>0.7370</td>
<td>0.7370</td>
<td>0</td>
<td>0</td>
<td>project</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
</tr>
<tr>
<td>determinist</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
</tr>
<tr>
<td>domin</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>robust</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
</tr>
<tr>
<td>estim</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>scale</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>experi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>separ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
</tr>
<tr>
<td>factor</td>
<td>0.7370</td>
<td>0.7370</td>
<td>0</td>
<td>0.7370</td>
<td>0</td>
<td>set</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fast</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>social</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>footprint</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>space</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fusion</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>tensor</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gener</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>text</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
</tr>
<tr>
<td>identifi</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>toolbox</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
</tr>
<tr>
<td>increment</td>
<td>0</td>
<td>0</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>user</td>
<td>2.3219</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>larg</td>
<td>1.3219</td>
<td>0</td>
<td>1.3219</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

✓ tf-idf
✓ Stemming
Text Data

... for text mining tasks

Retrieval
Text Data

... for text mining tasks

Retrieval

Clustering
Text Data

... for text mining tasks

- Retrieval
- Clustering
- Classification
Outline

1. Introduction to Text Data

2. The Text-to-Matrix Generator
What is TMG:

- Toolbox developed in University of Patras for text mining tasks over document collections
- Educational and Research tool

TMG: A MATLAB Toolbox for Generating Term-Document Matrices from Text Collections

(ZG06b)
What is TMG:

- Toolbox developed in University of Patras for text mining tasks over document collections
- Educational and Research tool

Implementation:

- over 17,000 lines of \texttt{matlab} and \texttt{perl}
- takes advantage from sparse technology provided by MATLAB
- first version by Zeimpekis (’06)
Six basic modules:

1. Indexing
2. Dimensionality Reduction
3. Non-Negative Matrix Factorizations
4. Retrieval
5. Clustering
6. Classification
How can I find TMG?

Free under request from:
http://scgroup20.ceid.upatras.gr:8000/tmg/

More than 4000 requests worldwide . . .

Part I

Introduction in version 6.0R7
Outline

1. Indexing Module
2. Dimensionality Reduction and Nonnegative Matrix Factorizations Modules
3. Retrieval Module
4. Clustering Module
5. Classification Module
6. Conclusions
Purpose

Document Collection

Term-by-Document Matrix

Graphical User Interface
Generate, Update and Downdate Term-by-Document Matrices II

Procedure
Generate, Update and Downdate Term-by-Document Matrices III

Supported non-ASCII formats

<table>
<thead>
<tr>
<th></th>
<th>ver.5.0R6</th>
<th>Filter ver.5.0R6</th>
<th>ver. 6.0R7</th>
<th>Filter ver. 6.0R7</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>TIKA</td>
</tr>
<tr>
<td>docx</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>htm</td>
<td>✓</td>
<td>strip_html</td>
<td>✓</td>
<td>strip_html</td>
</tr>
<tr>
<td>html</td>
<td>✓</td>
<td>strip_html</td>
<td>✓</td>
<td>TIKA</td>
</tr>
<tr>
<td>odt</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>TIKA</td>
</tr>
<tr>
<td>pdf</td>
<td>✓</td>
<td>ps2ascii</td>
<td>✓</td>
<td>ps2ascii</td>
</tr>
<tr>
<td>ps</td>
<td>✓</td>
<td>ps2ascii</td>
<td>✓</td>
<td>ps2ascii</td>
</tr>
<tr>
<td>rtf</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>TIKA</td>
</tr>
<tr>
<td>tex</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>Untex</td>
</tr>
</tbody>
</table>

Update

Update the TDM by inserting new documents

Downdate

Downdate the TDM by extracting useless documents
Outline

1. Indexing Module

2. Dimensionality Reduction and Nonnegative Matrix Factorizations Modules

3. Retrieval Module

4. Clustering Module

5. Classification Module

6. Conclusions
Dimensionality Reduction

Purpose

- Handling High Dimensional Data
  - Economical representation

- Reducing noise
  - Better semantic representation

Graphical User Interface

Image of a Graphical User Interface for a Text to Term-Document Matrix (tdm) Generator.
Dimensionality Reduction II

Available Methods

1. Singular Value Decomposition (SVD)
   - MATLAB svds
   - PROPACK svd (Larsen (Lar))

2. Centroids Method (CM) (Park, Jeon & Rosen (PJR03))

3. Semidiscrete Decomposition (SDD) (Kolda & O’Leary (KO00))

4. Clustered LSI (CLSI) (Zeimpekis & Gallopoulos (ZG05; ZG06a))

5. Sparse Pivoted QR Decomposition (SPQR) (Berry, Pulatova & Stewart (BPS05))

6. Principal Component Analysis (PCA)

SDD and SPQR call routines available from Netlib(TOMS)
Nonnegative Matrix Factorizations (NMF) I

**Purpose**

Factorizations on Nonnegative Matrices

⇓

Nonegative Factors

⇓

Preserving non-negativity

⇓

Better semantic representation

✓ Final results depend on initialization

✓ Resulting factors can be refined

**Graphical User Interface**

![Graphical User Interface](image-url)
## Initialization Techniques

1. Random Initialization
2. **Nonnegative Double SVD** NNDSVD (Boutsidis & Gallopoulos [BG08])
3. **Block Nonnegative Double SVD** (Zeimpekis & Gallopoulos [ZG08])
4. **Bisecting Nonnegative Double SVD** (Zeimpekis & Gallopoulos [ZG08])
5. By Clustering (Wild, Curry, Dougherty [WCD04])

## Factors Refinement

1. **Multiplicative Update Algorithm** (Lee & Seung [LS01])
2. Alternating Non-Negative-Constrained Least Squares (NMF/ANLS) (Kim & Park [KH08])

---

**NNDSVD** uses prepared implementation

**NMF/ANLS** uses prepared implementation
Outline

1. Indexing Module
2. Dimensionality Reduction and Nonnegative Matrix Factorizations Modules
3. Retrieval Module
4. Clustering Module
5. Classification Module
6. Conclusions
Purpose

Queries over a dataset

Retrieval I

Graphical User Interface

Retrieve all relevant documents via a HTML response
Available Methods

1. Vector Space Model (VSM) (Salton, Wong, & Yang (SWY75))
2. Latent Semantic Analysis (LSA) (Berry et al. (BDJ99; Dee+90))

**LSA can be combined with any DR or NMF technique**
Outline

1. Indexing Module
2. Dimensionality Reduction and Nonnegative Matrix Factorizations Modules
3. Retrieval Module
4. Clustering Module
5. Classification Module
6. Conclusions
Purpose

Collection of documents as a TDM

Clusters of related documents
Available Methods

1. Euclidean k-means
2. Spherical k-means (DM01)
3. Principal Direction Divisive Partitioning (PDDP) (Boley (Bol97))
4. PDDP (l) (Zeimpekis & Gallopoulos (ZG03))
5. PDDP (l) with some hybrid variants of PDDP and kmeans (Zeimpekis & Gallopoulos (ZG03))

PDDP(l) Variants

✓ Split with k-means
✓ Optimal Split
✓ Optimal Split with k-means
✓ Optimal Split on Projections
Classification I

Purpose

Collection of documents as training

TDM

+ List of training labels

↓

Assign new documents to related classes (labels)

Graphical User Interface
Available Methods

1. k Nearest Neighbours (\textit{knn})
2. Rocchio
3. Linear Least Squares Fit (\textit{LLSF}) (Yang & Chute (YC92))

✓ Combination with CLSI, CM and SVD DR techniques
✓ Implementations for multilabel and singlelabel collections
Outline

1 Indexing Module

2 Dimensionality Reduction and Nonnegative Matrix Factorizations Modules

3 Retrieval Module

4 Clustering Module

5 Classification Module

6 Conclusions
Goal:

😊 Make TMG more user friendly
Summary

Goal:

😊 Make TMG more user friendly

Work in Progress:

- Smarter parsing → boost parsing time
- Increase the degrees of freedom during parsing phase (e.g. stoplist, incorporation of new filters)
- Manual writing using MATLAB publish
- New stemming algorithms (e.g. greek stemmer)
- GUIs makeover
- Incorporation of new capabilities (e.g. WordNet, Wordle)
Thank you!
Questions ?


(Bol97) D. Boley. “Principal Direction Divisive Partitioning”. In: Data Mining and Knowledge Discovery 2 (1997), pp. 325-344.


(DM01) I. S. Dhillon and D. S. Modha. “Concept decompositions for large sparse text data using clustering”. In: Machine Learning 42.1 (2001), pp. 143-175.


(Lar) R.M. Larsen. PROPACK: A software package for the symmetric eigenvalue problem and singular value problems on Lanczos and Lanczos bidiagonalization with partial reorthogonalization.


